Appendix A. Technical Notes

Appendix A. Technical Notes[1]

The data on doctoral scientists and engineers contained in this report come from the 1997 Survey of Doctorate Recipients (SDR). The SDR is a longitudinal panel survey of individuals who have received their doctorates mainly in the sciences or engineering fields. Since the 1970s, this study has been conducted every two years for the National Science Foundation (NSF) and other federal sponsors.[2]

The National Opinion Research Center conducted the SDR for the first time in 1997. Data collected in the SDR are part of the Scientists and Engineers Statistical Data System (SESTAT) surveys that are sponsored and maintained by the NSF. Additional data on education and demographic information come from the Doctorate Records File (DRF), which contains data from an ongoing census of all research doctorates earned in the United States since 1920.

The Sampling Frame and Target Population

The sampling frame for the 1997 SDR was compiled from the DRF to include individuals who:

1. had earned a doctoral degree from a U.S. college or university in a science or engineering field;[3]
2. were U.S. citizens, or, if non-U.S. citizens, indicated they had plans to remain in the United States after degree award; and
3. were under 76 years of age.

The 1997 SDR frame consisted of the 1995 SDR sample supplemented with graduates who had earned their degrees since the 1995 survey and who met the conditions listed above. Those who were carried over from 1995 but had attained the age of 76 (or died) were deleted from the frame.

The survey had two additional eligibility criteria for the survey target population. The sampled member must be a resident in the United States and not institutionalized as of the survey reference date.

Sample Design

In 1997, the SDR sample size was 54,103. The total sample was selected from 2 groups:

1. 1995 sample members who were still eligible in 1997, and
2. a sample of the 1995-96 graduating cohort.

Group 2 cases were oversampled in 1997 to obtain more precise estimates on the recent doctorates data. A maintenance cut was done to the sample to keep the sample size of the Group 1 cases roughly the same as it was in 1995.

The basic sampling design was a stratified design where strata were defined by 15 broad fields of study, 2 genders, and an 8-category "group" variable combining race/ethnicity, handicap status, and citizenship status. As in the prior years, the goals were to maintain a fairly constant sample size and to equalize probabilities of selection to the extent possible. The primary changes for 1997 were an oversample of the 1995-96 cohort, and a slight redefinition of strata by field of study. The stratification variables were the same, but the classifications for field of study were revised in 1997. Humanities graduates were interviewed in 1995, but not in 1997.

The overall sampling rate was about 1 in 12 (8.5 percent) in the 1997 SDR, applied to an estimated population of 632,800. However, sampling rates varied considerably within and between the strata. These differences resulted from oversampling to provide a useful sample size for the recent doctorate cohorts, women, minority groups and other groups of special interest, and the accumulation of sample size adjustments over the years.

Survey Content

The 1997 SDR retained questionnaire design changes that were implemented in 1993. In addition to a large set of core data items that are conveyed from year to year, the 1997 questionnaire included new questions covering several areas of interest. The 1995 modules on the work history and postdocs were dropped

[1] The discussions presented here are partly from The Methodological Report of the 1997 Survey of Doctorate Recipients (NORC, March 1999).

[2] In 1997, the National Institutes of Health co-sponsored the SDR with NSF. In previous rounds, the Department of Energy and the National Endowment for the Humanities co-sponsored the survey. Until 1995, the SDR was conducted by the National Research Council (NRC).

[3] See appendix B for a list of the specialties included in the 1997 SDR sampling frame.

and a new module on the recent doctorates was added in 1997. Also a new question was asked of the respondents to classify employer's main business in addition to a series of questions on temporary or alternative work arrangements, job security concerns, job satisfaction, and household income.

Data Collection

The 1997 SDR data collection consisted of two phases: a self-administered mail survey, followed by computer assisted telephone interviewing (CATI) of a sample of the nonrespondents to the mail survey. The mail survey consisted of an advance letter and the several waves of a personalized mailing package, with a reminder postcard between the 1st and 2nd questionnaire mailing. The advance letter was sent in May 1997, followed by the 1st mailing in early June. The second mailing was sent in August 1997. To increase the mail response rate, an additional follow-up mailing occurred via Federal Express. The CATI follow-up ended in March 1998.

Response Rates

The overall unweighted response rate for the 1997 SDR was 85 percent. The response to the mail phase of the survey was about 55 percent. The overall weighted response rate was about 78 percent (weighted response divided by the weighted sample cases.)

Data Preparation

Data preparation for the 1997 SDR included pre-data entry edit, data entry, coding, telephone call backs for critical items and sample verification, post-data entry editing and data review, and imputation. As completed survey mail questionnaires were received, they were logged and transferred to the pre-data entry editing at NORC for processing.

The data from the questionnaire were keyed into the database in a process known as CADE (Computer-Assisted Data Entry). The data entry program, SurveyCraft, contained a full complement of range, consistency, skip error checks to prevent entry errors and inconsistent answers. Three on-line coding programs were tied into the SDR CADE program to ease data entry of special codes: IPEDS for educational institutions, Federal Information Processing Standards (FIPS) for U.S. states and foreign countries, and Primary Field of Study/Education. Consistency checks were also built into the CATI program along with the skip patterns. Some consistency checks were performed on a number of variables prior to the merge of the CADE and CATI data files to ensure complete compatibility. Computer checks also flagged the cases with missing key items (employment status, occupation, birthdate, etc.) and the telephone call-backs were made to obtain the response; otherwise they were considered as incomplete responses.

A detailed edit specification was developed from the SESTAT surveys edit guideline to perform further computer editing of multiple values to "Mark One" questions, skip errors, range errors, inter-item inconsistencies, cross year inconsistencies. "Other Specify" responses were coded using the SESTAT coding guidelines and respondents' occupational data was reviewed along with other work-related data from the questionnaire to "correct" known respondent self-reporting problems to obtain the "best" occupation codes.

Basic frequency distributions of all survey items showed item nonresponse rates to be generally less than 3 percent. Nonresponse to a few questions deemed somewhat sensitive, such as annual salary or household income, was around 6.5 percent. To compensate for the item nonresponse, data not reported by the respondents, as well as response of "refused" or "don't know" were imputed. Two imputation methods were used: (1) logical imputation, and (2) hot deck imputation. For logical imputation, either the respondent's answers to related questions determined what the missing value had to be, or the respondent's answer to the same question in the prior survey round substituted for the missing value. The latter approach of using the historical data is often called "cold deck" imputation. Cold deck imputation is useful for variables that are static, such as place of birth or gender. When logical imputation was used, it was employed before hot deck imputation.

In hot deck imputation, a donor case is selected from the current round of respondents by matching on related variables. The donor case's response is used as a proxy for the recipient's missing variable. Hot deck imputation is the method of choice for variables that may change over time, such as employment characteristics. Hot deck is preferable to model-based imputation in this application because it easily preserves correlation among variables and maintains the valid response rages for categorical variables.

Imputation was done in a specified sequence, with key auxiliary variables being imputed first. After the key variables were imputed, variables were imputed by questionnaire section. Within a section, variables were imputed more or less in questionnaire order, with certain exceptions. Questions used to drive skip patterns were imputed before questions affected by the skip driver. Questions new to this round were imputed last within a section. Where logical, groups of companion variables were imputed together (such as the various reasons for working outside the Ph.D. field).

Weighting and Estimation

To enable weighted analyses of the 1997 SDR data, a sample weight was calculated for every person in the sample. The primary purpose of the weights is to create representative estimates by adjusting for unequal probabilities of selection. The second purpose is to adjust for the effects of nonresponse. Informally, a sampling weight approximates the number of persons in the Ph.D. population that a sampled person represents.

The weights were calculated in several stages. The first stage was the calculation of base weights that account for the sample design. A base weight for a respondent is the reciprocal of the probability of selection. The revised base weights ranged from 1.0 to 112.008 with a median value of 11.442. The sum of the revised weights, 632,789, is also an estimate of the frame size. Base weights varied within cells because different sampling rates were used depending on the year of selection and the stratification in effect at that time.

The next stage was to construct a combined weight, which took into account the subsampling of nonrespondents at the CATI phase. All respondents received a combined weight, which for mail respondents was equal to the sample weight and for CATI respondents was a combination of their original sample weight and their CATI subsample weight. The final stage was to adjust the sampling weights for unit nonresponse. (Unit nonresponse occurs when the sample member refuses to participate or cannot be located.) This was done in a group of nonresponse adjustment cells created using poststratification.

Within each nonresponse adjustment cell, a weighted nonresponse rate, which took into account both mail and CATI nonresponse, was calculated. The nonresponse adjustment factor was the inverse of this weighted response rate. The initial set of nonresponse adjustment factors was examined and, under certain conditions, some of the cells were collapsed if use of the adjustment factor would create excessive variance. The final weights for respondents were calculated by multiplying their respective combined weights by the nonresponse adjustment factor. In data analysis, population estimates are made by summing the final weights of all respondents who possess a particular characteristic.

Reliability

Because the estimates produced from this survey are based on a sample, they may vary from those that would have been obtained if all members of the target population had been surveyed (using the same questionnaire and data collection methods). Two types of error are possible when population estimates are derived from measures of a sample: nonsampling error and sampling error. By looking at these errors, it is possible to estimate the accuracy and precision of the survey results.

Sampling error is the variation that occurs by chance because a sample, rather than the entire population, is surveyed. The particular sample that was used to estimate the 1997 population of science and engineering doctorates in the United States was one of a large number of samples that could have been selected using the same sample design and size. Estimates based on each of these samples would have differed.

Sampling errors were developed using a generalized variance procedure in order to provide approximate sampling errors that would be applicable to a wide variety of items. As a result, these sampling errors provide an indication of the order of magnitude of a sampling error rather than a precise sampling error for any specific item. This method first computes the variances associated with selected variables for certain subsets of the sample. The variances of the selected variables were computed using SUDAAN software and the Taylor series approximation method, which can incorporate finite correction factors. The finite correction factors are important for the SDR sample design where some strata had high sampling fractions.

The estimated variances for the selected variables were used to estimate regression coefficients for use in generalized variance functions that estimate the standard errors associated with a broader range of totals and percentages. For each of the demographic groups and fields of study shown in Appendix D, 31 models from the variables listed above were combined into a nonlinear regression to fit a predictive model for standard errors, as described below.

Appendix table D shows model parameters, *a* and *b*, that can be used to approximate standard errors for the S&E doctoral population overall, for broad field groupings used by NSF, and for selected subgroups of analytic interest.[4] Let *x* denote the estimated total for which a standard error is desired. The standard error can be approximated using the appropriate values of *a* and *b* along with the following formula for standard errors of totals:

$$Sx = [ax^2 + bx]^{1/2}$$

Percentages are another type of estimate for which standard errors may be desired. The standard error of a percentage may be approximated using the formula:

$$Sp = p[b((1/x)-(1/y))]^{1/2}$$

where p equals the percentage possessing the specific characteristic and x and y represents the numerator and denominator, respectfully, of the ratio that yields the observed percentage.

In addition to sampling error, data are subject to nonsampling error, which can arise at many points in the survey process. Sources of nonsampling error takes many different forms: (1) nonresponse bias, which arises when the characteristics between individuals who do not respond to a survey differ significantly from those who do; (2) measurement error, which arises when we are not able to precisely measure the variables of interest; (3) coverage error, which arises when some members of the target population are not identified and thus do not have a chance to be selected for the sample; (4) processing error, which can arise at the point of data editing, coding or key entry. These sources of error are much harder to estimate than sampling errors.

Important Notes on the Tables

Please note several changes that were made in the 1997 tables from 1993 and 1995 reports:

1. **Doctorate field groups** were changed as follows:
   - Health sciences is now shown separately from the biological sciences (characteristics between these two field are deemed to be too different to be shown combined);
   - Other physical sciences, including earth sciences, were combined with geology and oceanography to form a new combined group, earth/atmospheric/ocean sciences (individual field counts are too small thus the meaningful groups are combined together);
   - Anthropology is separated from sociology and is combined with other social sciences;
   - Psychology is now shown separately from the social sciences (characteristics between psychology and other social sciences are deemed to be too different to be shown combined);
   - Industrial engineering is combined with other engineering (number was getting too small); materials/metallurgical engineering is now shown separately; and
   - Computer/information sciences and mathematical sciences are now shown separately in all broad doctorate field tables (characteristics between these two fields are deemed to be too different to be shown combined).
2. **Occupation field groups** were changed as follows:
   - Psychologists and postsecondary teachers in psychology are shown separately from social sciences.
   - Computer/information scientists and mathematical scientists are now shown separately in all broad occupation tables.
3. Following **table number changes** occurred:

1993 and 1995 tables no.	1997 table no.
17	21
18	22
19	23
20	17
21	18

4. Because of the many redesign changes introduced to the 1993 SDR still retained in 1997, users are advised that the data in this report, as well as the in the 1993 or 1995 reports, are not strictly comparable with the SDR data published by NSF prior to 1993.

[4] The generalized error estimates in this report were based on a set of assumptions that did not appear to hold in the case of some small subpopulations. In such cases, the parameters listed for a higher-level field within a demographic group or a higher-level demographic group within a field were considered a useful substitute as a generalized error estimate.

The following notes will help facilitate the use of data in the detailed tables.

Field of doctorate is the field of degree as specified by the respondent in the Survey of Earned Doctorates at the time of degree conferral. (See appendix B for doctorate degree field.)

Occupation data were derived from responses to several questions on the type of work primarily performed by the respondent. The occupational classification of the respondent was based on his/her principal job held during the reference week— or last job held, if not employed on the reference week (questions A26 or A5). Also used in the occupational classification was a respondent-selected job code (questions A27 or A6).

Sector of employment was based on responses to questions A15 and A17. The category "universities and 4-year colleges" includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), university affiliated research institutions, and other type of institutions. "Private-for-Profit" includes self-employed in incorporated business.

Employer Location was based primarily on responses to question A11 on the location of the principal employer. Individuals not reporting place of employment were classified by their last mailing address.

Place of Birth categories were defined as follows:

U.S. = Fifty states plus the Virgin Islands, Panama Canal Zone, Puerto Rico, American Samoa, Trust Territory, and Guam

Europe = Albania, Armenia, Austria, Belarus, Bosnia-Herzegovina, Bulgaria, Czech Republic, Croatia, Estonia, Georgia, Greece, Hungary, Latvia, Lithuania, Poland, Romania, Russia, Slovakia,Ukraine, Federal Republic of Yugoslavia, Andorra, Belgium, France, Gibraltar, Luxembourg, Monaco, The Netherlands, Portugal, Spain, Switzerland, Germany, Italy, Liechtenstein, Malta, Denmark, England, Finland, Iceland, Northern Ireland, Republic off Ireland, Norway, Scotland, Sweden, Wales, Europe, not specified

Asia = Afghanistan,Bahrain,Bangladesh, Cyprus, India, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Nepal, Palestine, Saudi Arabia, Sri Lanka, Syria, Turkey, Cambodia, People's Republic of China, Philippines, Taiwan, China Unspecified, Hong Kong, Japan, Republic of Korea, Korea Unspecified, Laos, Malaysia, Singapore, Thailand, Democratic Republic of Vietnam, Republic of Vietnam, Asia, not specified

North America = Bermuda, Canada, Greenland, North America, not specified

Central America = Belize, Costa Rica, El Salvador, America Guatemala, Honduras, Mexico, Nicaragua, Panama, Central America, not specified

Caribbean = Barbados, Cuba, Dominican Republic, Haiti, Jamaica, Caribbean not specified

South America = Argentina, Bolivia, Brazil, Chile, Columbia, Ecuador, French Guinea,Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela, South America, not specified

Africa = Algeria, Egypt, Ethiopia, Ghana, Kenya, Libya, Morocco, Nigeria, South Africa, Sudan, Africa, not specified

Oceania = Australia, Indonesia, New Zealand, Oceania, not specified

Primary work activity was determined from responses to question A38. "Development" includes the development of equipment, products, and systems. "Design" includes the design of equipment, processes, and models.

Federal support was determined from responses to questions A46 and A47.

Faculty Rank/Tenure status was obtained from the response to questions A18 and A19.

Race/ethnicity categories of white, black, Asian/Pacific Islander and American Indian/Alaskan Native refer to non-Hispanic individuals only.

Citizenship status category of Non-U.S., temporary resident does not include individuals who, at the time they received their doctorate, expressed plans to leave the U.S. These individuals were excluded from the sampling frame.

Salary data were derived from responses to question A43, in which information was requested regarding annual salary before deductions for the principal job held during April 1997, excluding income from bonuses, overtime, and summer teaching/research. Salaries reported are median annual salaries, rounded to the nearest $100 and computed for full-time employed scientists and engineers. For individuals employed by educational institutions, no accommodation was made to convert academic-year salaries to calendar-year salaries. Users are advised that due to a wording change in the salary question since 1993, the 1997 salary data are not strictly comparable with 1993 salary data.

Labor force participation rate. The labor force is defined as those employed (E) plus those unemployed (U—i.e., those not-employed persons actively seeking work). Population (P) is defined as all S&E doctorate holders under age 76, residing in U.S. during the week of April 15, 1997, who earned their doctorate from U.S. institutions. The labor force participation rate (R_{LF}) is the ratio of the labor force to the population (P).

$$R_{LF} = (E+U) / P$$

Unemployment rate. The unemployment rate (R_U) is the ratio of those who are unemployed but seeking employment (U) to the total labor force (E+U).

$$R_{LF} = U / (E+U)$$

Involuntarily out-of-field rate. The S&E involuntarily out-of-field rate is the percent of employed individuals who reported they were either:

- working part-time exclusively because suitable full-time work was not available; and/or

- working in an area not related to the first doctoral degree (in their principal job) at least partially because suitable work in the field was not available.